# SUPPLEMENTARY MATERIAL

# To randomize, or not to randomize? That is the question: a meta-analytic methodology for determining the context-specific value of randomization.

March 11, 2019

# S1 Supplementary Analyses and Figures

## S1.1 Figures and tables referenced in the main text

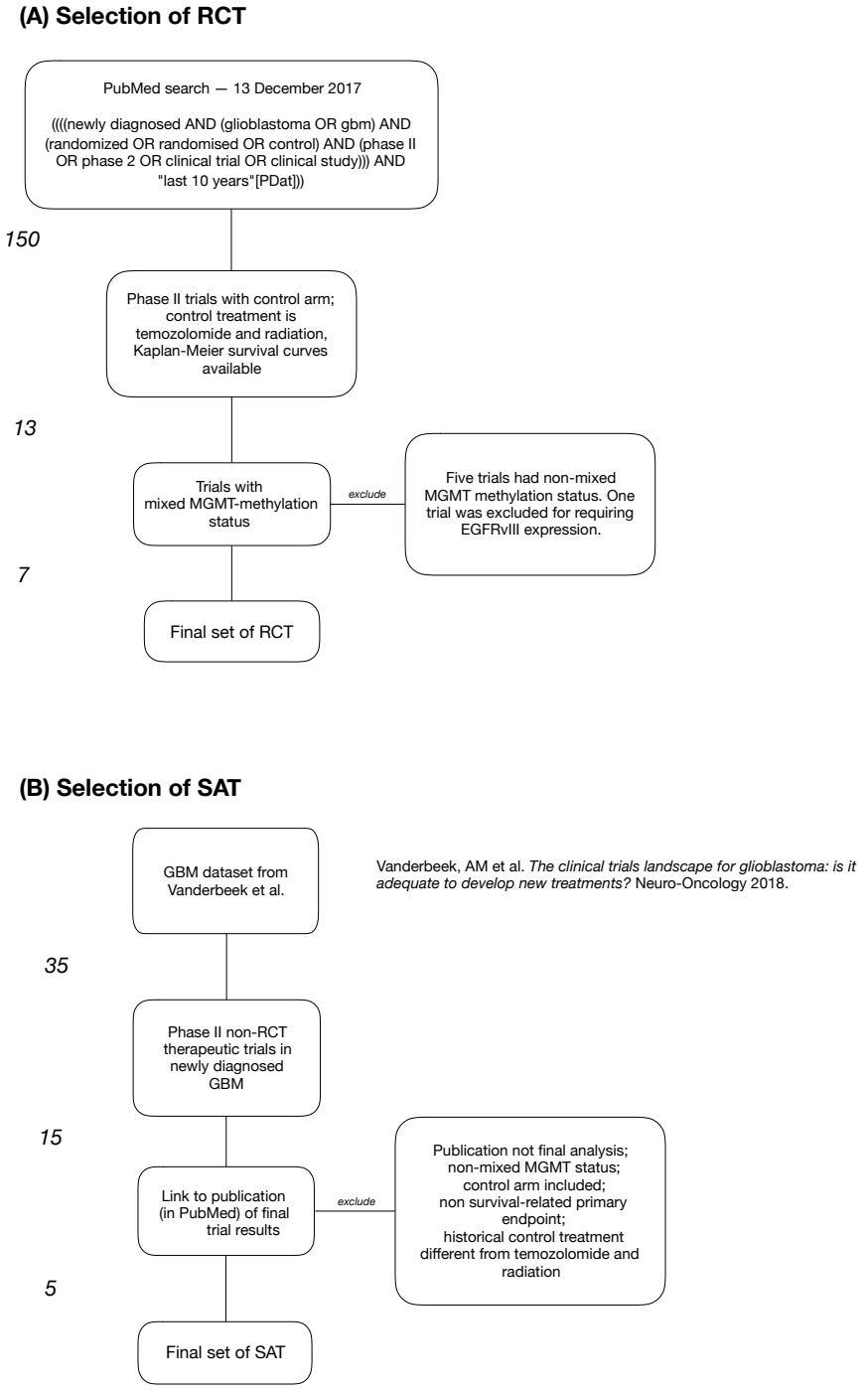


Figure S1: Systematic literature review and selection of RCTs and SATs for our meta-analysis.

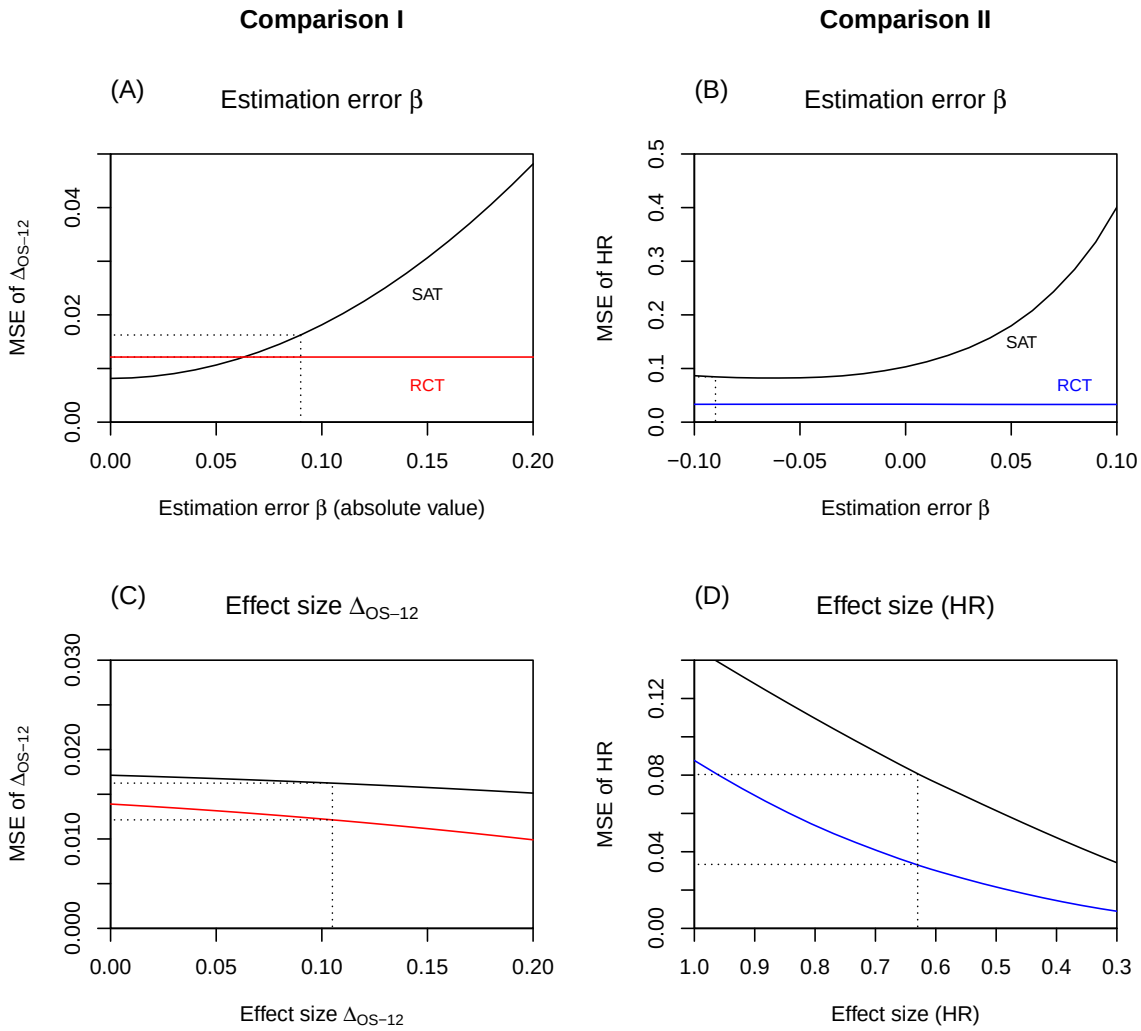


Figure S2: MSE of the estimated treatment effect over a range of values for the estimation error $\beta$ (Panels A and B) and treatment effect $\Delta$ (Panels C and D). The SAT design with OS-12 endpoint (black curve), RCT with OS-12 (red curve), and RCT with OS endpoint (blue curve) have a total trial sample size of 60 patients. The average OS-12 response rate of the SOC and variability of the SOC's response rate across trials is fixed at the estimated values $\widehat{p}_0 = 0.7$ and $\widehat{\sigma}_{SOC} = 0.075$. Horizontal grey lines indicate the MSE at the observed parameters $\widehat{\beta}$ and $\Delta$.

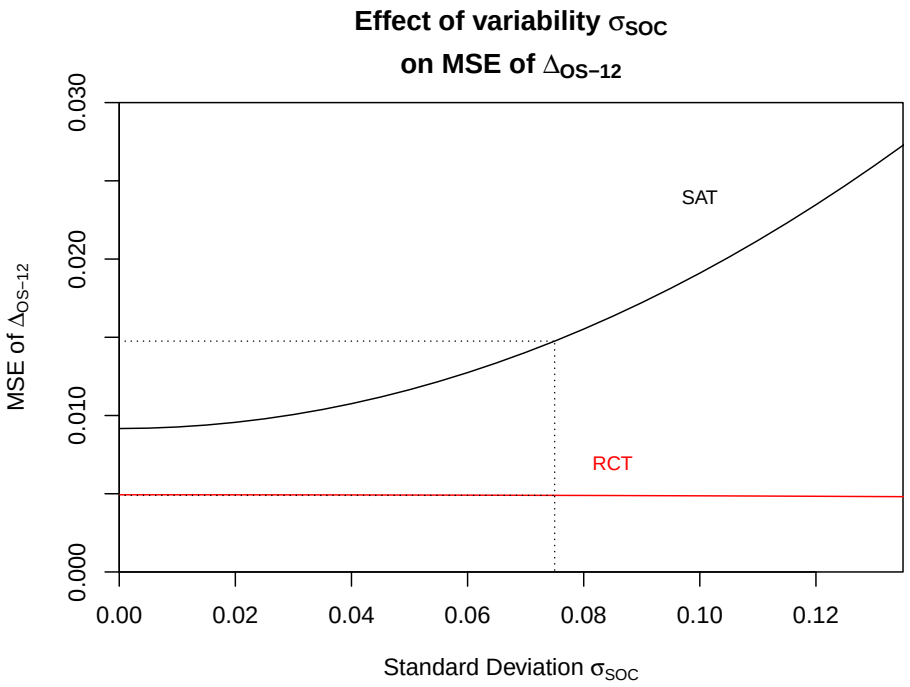


Figure S3: MSE of the estimated treatment effect for a range of variability values $\sigma_{SOC}$ of the SOC's OS-12 across trials for a SAT design (black curve) and RCT design (red curve) both with OS-12 outcomes and 150 patients. The average OS-12 response rate of the SOC's response rate across trials is fixed at the estimated values $\widehat{p}_0 = 0.7$ and $\Delta = 0.1$. Compared to Figure 1C in the main paper, the RCT has a smaller MSE of $\Delta$ across all possible values of $\sigma_{SOC}$.

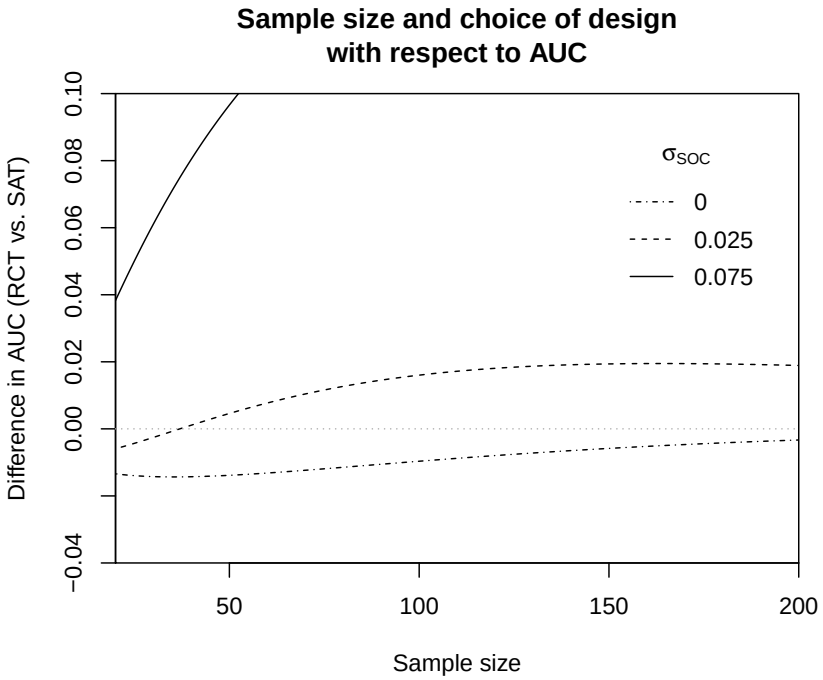


Figure S4: Difference in AUC of the RCT design with OS endpoint compared to the SAT design with OS-12 endpoint.

## S1.2 Designs with larger sample sizes

In the main paper we compared RCT and SAT designs for an overall sample size of 60 patients. We show an additional analysis which is identical to comparison II in the main text, except we assume an overall sample size of 150 patients for both the RCT and SAT. As Figure **??** shows, with larger sample size, the RCT is superior to the SAT with AUC of 0.97 compared to 0.82 for the SAT. Additionally, the RCT has 94% power and controls the type I error at 10%, compared to the SAT with 96% power and a high 69% type I error level. Additionally, the SAT has a higher MSE than the RCT (0.073 compared to 0.012).

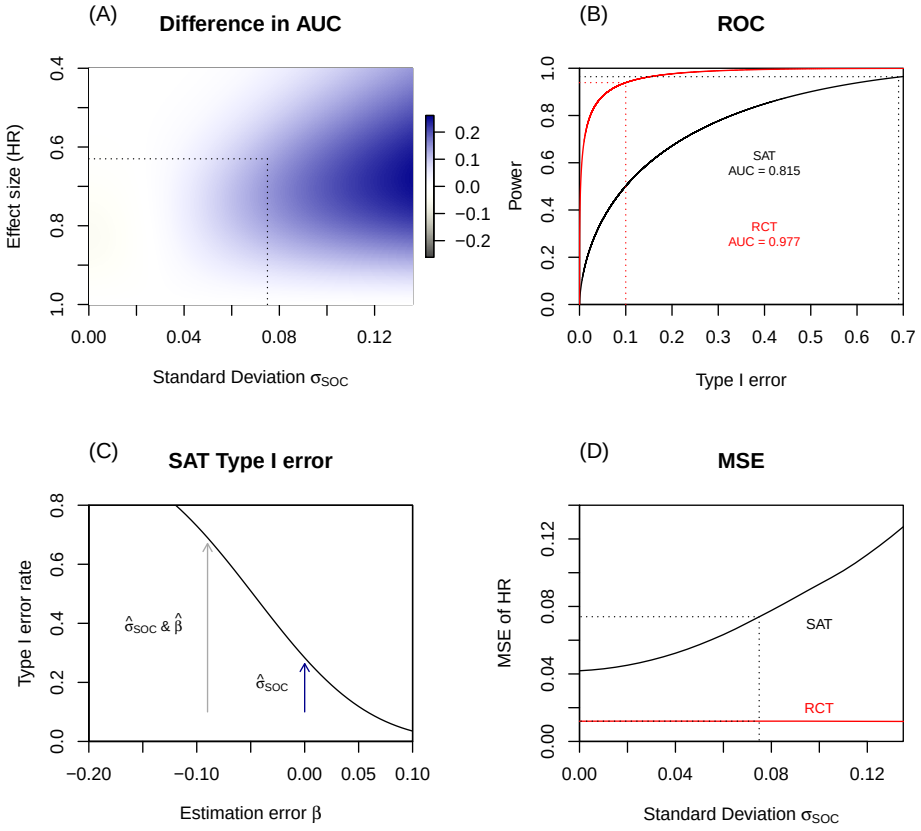


Figure S5: The figure shows the difference in AUC (Panel A), ROC curves (Panel B), type I/II error rates (Panel B and C), and the MSE of the estimated treatment effect (Panel D) for a SAT and RCT with 150 patients and OS-related outcomes. The vertical and horizontal lines in Panel (A), (C), and (D) indicate the estimated values of $\sigma_{SOC}, \beta, \Delta_{OS}$ at which we compute the difference in AUC, type I and II error rates, and the MSE of the estimated treatment effect.

## S2 Metrics of Design Comparison

We use the AUC, type I/II error rates, and the mean squared error (MSE) of the estimated treatment effect to compare SAT and RCT designs. For SAT and RCT designs, we let $\widehat{\Delta}$ indicate the estimated treatment effect of the experimental arm compared to the SOC based on the data generated by the trial. The MSE is the average squared difference between the treatment effect estimate $\widehat{\Delta}$ and the treatment effect $\Delta$,

$$MSE = E[(\widehat{\Delta} - \Delta)^2] = \text{Var}(\widehat{\Delta}) + (E[\widehat{\Delta}] - \Delta)^2. \tag{S1}$$

Each (SAT and RCT) design uses a statistics $Z$ to test efficacy ($H_0 : \Delta = \Delta_0$ vs. $H_A : \Delta > \Delta_0$). The statistics are approximately normally distributed with mean $\mu_\Delta$ and variance $\sigma^2_\Delta$. Designs may use various statistics, for instance the SAT design and the RCT design with binary outcomes use z-statistics for one-sample (SAT) and two-sample (RCT) test for proportions. The test rejects $H_0$ if $Z \geq z_{1-\alpha}$ where $z_{1-\alpha}$ is a threshold to control a targeted type I error rate at level $\alpha$. For a fixed design, consider independent variables $Z_0 \sim N(\mu_{\Delta_0}, \sigma^2_{\Delta_0})$ and $Z_A \sim N(\mu_{\Delta_A}, \sigma^2_{\Delta_A})$ having (approximately/asymptotically) the same distributions as the test statistic under $H_0$ and $H_A$ (assuming $\Delta = \Delta_A$ under $H_A$). The AUC index [**?** ] is defined as

$$AUC = Pr\Big(Z_A > Z_0\Big) = 1 - \Phi\bigg( - \frac{\mu_{\Delta_A} - \mu_{\Delta_0}}{\sqrt{\sigma^2_{\Delta_A} + \sigma^2_{\Delta_0}}}\bigg), \tag{S2}$$

where $\Phi$ is the cumulative distribution function of the standard normal distribution. The power function (type I error under $H_0$ and power under $H_A$), at a threshold value $z$ equals $P(z, \Delta) = 1 - \Phi\Big(\frac{z-\mu_\Delta}{\sigma_\Delta}\Big)$.

We consider a SAT and RCT design, both with a binary endpoints, and a RCT design with time-to-event outcome. All three designs have an identical overall sample size of $n$ patients. The RCT randomizes patients in equal proportions (1:1) to the control and experimental arms.

For designs with binary endpoints, $p_{k,i}$ indicates the unknown response probability for the control ($k = 0$) and experimental ($k = E$) arm in trial $i$, while $\widehat{p}_{k,i}$ indicates estimated response rates. We assume that $p_{E,i} = p_{0,i} + \Delta$ and $p_{0,i} = p_0 + \epsilon_i$, where $E[\epsilon_i] = 0, \text{Var}[\epsilon_i] = \sigma^2_{SOC}$.

### Single Arm Trial design

The SAT design compares the estimated response rate $\widehat{p}_{E,i}$, to the threshold $p_{0,SAT}$, which is selected based on past trial data. The variable $\frac{\widehat{p}_{E,i} - p_0 + \Delta}{\sqrt{(p_0+\Delta)(1-p_0-\Delta)/n + \sigma^2_{SOC}}}$ is approximately standard normal distributed for large $n$, and the design uses the one-sample z-test for proportions with statistics

$$Z_{SAT} = \frac{\widehat{p}_{E,i} - p_{0,SAT}}{\sqrt{p_{0,SAT}(1 - p_{0,SAT})/n}}.$$

The test rejects the null hypothesis $H_0 : \Delta = p_{E,i} - p_{0,i} = 0$ when $Z_{SAT}$ exceeds $z_{1-\alpha}$. The statistics $Z_{SAT}$ is approximately normal distributed with mean $\mu_\Delta = \frac{\Delta - \beta}{\sqrt{p_{0,SAT}(1-p_{0,SAT})/n}}$ and variance $\sigma^2_\Delta =$

$\frac{(p_0+\Delta)(1-p_0-\Delta)/n+\sigma^2_{SOC}}{p_{0,SAT}(1-p_{0,SAT})/n}$ where $\beta$ indicates the estimation error $\beta = p_{0,SAT} - p_0$. The AUC and power function are then given by

$$AUC_{SAT} \approx 1 - \Phi\left(\frac{-\Delta}{\sqrt{2\sigma^2_{SOC} + \Big(p_0(1-p_0) + (p_0+\Delta)(1-p_0-\Delta)\Big)/n}}\right) \text{ and} \tag{S3}$$

$$P(z,\Delta) \approx 1 - \Phi\left(\frac{z}{\sigma_\Delta} - \frac{\Delta - \beta}{\sqrt{(p_0+\Delta)(1-p_0-\Delta)/n + \sigma^2_{SOC}}}\right). \tag{S4}$$

Whereas the MSE of the treatment effect equals

$$MSE(\Delta) = \sigma^2_{SOC}(1-1/n) + \beta^2 + ((p_0+\Delta)(1-p_0-\Delta)/n.$$

## Randomized Controlled Trial designs

The RCT with binary endpoints contrasts the observed response rate of the experimental with the control rate using a two-sample test for proportions with statistics $Z_{RCT} = \frac{\widehat{p}_{E,i} - \widehat{p}_{0,i}}{\sqrt{Var(\widehat{p}_{E,i} - \widehat{p}_{C,i})}}$ that follows approximately a normal distribution with mean $\mu_\Delta = \frac{\Delta}{\sqrt{(p_0(1-p_0)+(p_0+\Delta)(1-p_0-\Delta))/(n/2)}}$ and unite variance. The AUC and power function of the RCT design equal $AUC \approx 1 - \Phi(-\mu_{\Delta_A}/\sqrt{2})$ and $P(z,\Delta) \approx 1 - \Phi(z - \mu_\Delta)$, while the MSE equals

$$MSE(\Delta) = \Big(p_0(1-p_0) + (p_0+\Delta)(1-p_0-\Delta) - 2\sigma^2_{0,SOC}\Big)/(n/2).$$

The RCT design with time-to-event endpoint and measures the treatment effect between the experimental arm and the SOC assuming a constant hazard ratio $\Delta = HR$ using a one-sided log-rank test, $H_0 : HR = 1$ vs $H_A : HR < 1$.

Under censoring completely at random, the log-rank statistics $Z$ follows approximately a normal distributed [**?** ] with unit variance and mean $\mu_{HR} = -\log(HR)\sqrt{nd/4}$, where $d$ is the proportion of patients that are expected to have an event by the end of the follow-up period of the trial. The power function and AUC of the design are given by

$$AUC_{RCT} \approx 1 - \Phi\left(\log(HR)\sqrt{nd/8}\right), \tag{S5}$$

$$P(z,HR) \approx \Phi(z + \log(HR)\sqrt{nd/4}). \tag{S6}$$

In our calculations, we assumed $d = 1$; based on the accrual rates and observed follow up period of the RCT trials in our meta-analysis, $d$ ranges between 0.75 and 0.95.

The value $d = 0.75$ translates into an AUC of 0.84 compared to 0.89 when $d = 1$, and 0.78 and 0.749 for the SAT and RCT with OS-12. Similarly, the power of of the RCT at $HR = 0.1$ equals 61%, 68% and 70% when $d = 0.75, 0.95$ or $d = 1$.

If we are interested in OS, then under the proportional hazard model $p_{k,i} = S_0^{HR}(t) = p_{0,i}^{HR}$ with

$t = 12$, the OS-12 treatment effect $\Delta_{OS12}$ translates into a hazard ratio $HR = \frac{\log(p_0 + \Delta_{OS-12})}{\log(p_0)}$.

When the target effect is measured as HR, the MSE for both the SAT (binary endpoint) and RCT with time-to-event endpoint is computed with the following code:

```
library(survival)
set.seed(1)

# Parameters
pc             = 0.7          #average SOC response
bias           = -0.09       #average bias
p.c.IG         = 0.7 + bias   #investigator's threshold
sd.c           = 0.075        #variability in SOC
Delta.binary   = 0.1          #expected effect size (binary)
Delta.hr       = 0.63            #expected HR
n              = 60           #total sample size
mc             = 10^4         #number of simulations
t.m            = 12           #time point for OS rate

# Simulate outcomes
p.c.mc    = rnorm(n=mc, mean=pc,  sd=sd.c)
p.c.mc    = p.c.mc[  p.c.mc<1 & p.c.mc>0  ]
h.c.mc    = -log(p.c.mc)/t.m
h.E.mc    = h.c.mc * Delta.hr
n.tr      = length(h.c.mc)

# SAT
p.E.mc    = p.c.mc + Delta.binary
p.E.mc    = p.E.mc[ p.E.mc>0 & p.E.mc<1 ]
HR.SAT    = log(rbinom(n=n.tr, size=n, prob=p.E.mc)/n) / log(p.c.IG)
MSE.SAT   = mean((HR.SAT - Delta.hr)^2)

# RCT
x         = rep(c(0,1), each = n/2)
c         = rep(1, n)
HR.RCT    = exp(sapply(1:n.tr, function(i)  {
  coxph(Surv(c(rexp(n/2, h.c.mc[i]), rexp(n/2, h.E.mc[i])), c)~x)$coef[1]
}))
MSE.RCT   =  mean((HR.RCT - Delta.hr)^2)

print(MSE.SAT)
print(MSE.RCT)
```